

Reweighted Data for Robust Probabilistic Models

Yixin Wang
Columbia University

Alp Kucukelbir
Columbia University

David M. Blei
Columbia University

June 14, 2016

Abstract

Probabilistic models analyze data by relying on a set of assumptions. When a model performs poorly, we challenge its assumptions. This approach has led to myriad hand-crafted robust models; they offer protection against small deviations from their assumptions. We propose a simple way to systematically mitigate mismatch of a large class of probabilistic models. The idea is to raise the likelihood of each observation to a weight. Inferring these weights allows a model to identify observations that match its assumptions; down-weighting others enables robust inference and improved predictive accuracy. We study four different forms of model mismatch, ranging from missing latent groups to structure misspecification. A Poisson factorization analysis of the MovieLens dataset shows the benefits of reweighting in a real data scenario.

1 Introduction

Probabilistic modeling is a powerful approach to discovering hidden patterns in modern data. We begin by expressing our assumptions about the class of patterns we expect to discover; this is how we design a probability model. We follow by inferring the posterior of the model; this is how we reason about the patterns we discover. Advances in automated inference ([Hoffman and Gelman, 2014](#); [Mansinghka et al., 2014](#); [Kucukelbir et al., 2016](#)) now enable easy development of new models for machine learning and artificial intelligence ([Ghahramani, 2015](#)).

What makes a model “good” or “bad”? We think of bad models as having bad assumptions. But this is not always the case. Consider data where a few observations are corrupted; they do not belong to the process we are modeling yet they appear in the data anyway. Imagine a movie recommendation scenario. A child logs in to her account and regularly watches popular animated films. One day, her parents log in and watch an obscure horror movie. This account now has a corrupted measurement: the horror movie. If a model were to perform well without this corrupted measurement, we would not consider it “bad”.

One strategy is to design new models that are less sensitive to corrupted data; for instance, we could replace a Gaussian likelihood with a heavier-tailed t distribution. This approach leads to robust models that mitigate this specific type of mismatch ([Huber, 2011](#); [Insua and Ruggeri, 2012](#)).

Yet, there are other types of model mismatch. For example, a mixture model may not have enough components to describe a dataset with many latent groups. Or, a regression model may have misspecified dependencies on its covariates. In this paper, we develop a way of systematically mitigating the gap between model assumptions and reality.

Main idea. We propose reweighted probabilistic models (RPMs). The idea is simple. Begin with a probabilistic model and adjust the contribution of each observation by raising each likelihood term to its own weight. Then infer these weights along with the latent variables of the original probability model. An RPM automatically identifies observations and hidden patterns that match its assumptions; observations that disagree with its assumptions get weighted down.

Consider a dataset of N independent observations $y = \{y_n\}_1^N$. The likelihood factorizes as a product $\prod_n \ell(y_n | \beta)$, where β is a set of latent variables. Posit a prior distribution $p_\beta(\beta)$.

Mathematically, the data reweighting approach follows these four steps.

1. Begin with a probabilistic model $p_\beta(\beta) \prod_1^N \ell(y_n | \beta)$.
2. Raise each likelihood to a positive weight w_n ; the model becomes $p_\beta(\beta) \prod_1^N \ell(y_n | \beta)^{w_n}$.
3. Choose a prior on the latent weights $p_w(w)$, where $w = (w_1, \dots, w_N)$.
4. Infer the posterior $p(\beta, w | y)$.

The latent weights w allow an RPM to explore which observations match its assumptions and which do not. The prior $p_w(w)$ plays an important role during this task. We explore three options, each encoding a different *attitude* towards the original model’s assumptions.

Bank of Beta priors. This option constrains all weights to a maximum value of one. Thus, observations can only be down-weighted, which increases posterior uncertainty. Overfitting cannot occur; this is the most cautious attitude.

Scaled Dirichlet prior. This option ensures the sum of the weights equals the number of observations N . However, each observation can be up- or down-weighted. Depending on the parameter of the Dirichlet prior, overfitting is possible; this is a moderate attitude.

Bank of Gamma priors. This option does not constrain the weights. Thus, observations can be arbitrarily up- or down-weighted. Overfitting is a real threat here; this is only for the intrepid. (We use it for theoretical analysis only.)

Section 2 presents these options in full detail, along with theory and intuition. Inferring the RPM posterior $p(\beta, w | y)$ may seem daunting. This is where automated inference algorithms shine. In section 3, we study four models under various forms of model mismatch. In each case, we use Stan (Carpenter et al., 2015), a probabilistic programming system, for inference. Section 4 presents a recommendation system example, where we identify atypical film enthusiasts in the Movielens 1M dataset.

Related work. This work draws on two themes around robust modeling. The first is a rich body of work on robust statistics and machine learning (Provost and Fawcett, 2001; Song et al., 2002; Yu et al., 2012; McWilliams et al., 2014; Feng et al., 2014; Shafieezadeh-Abadeh et al., 2015). These developments focus on making specific models more robust to imprecise measurements. One strategy appears popular: localization.

To localize a probabilistic model, allow each likelihood to depend on its own “copy” of the latent variable β_n . This transforms the model into

$$p(y, \beta, \alpha) = p_\alpha(\alpha) \prod_n^N \ell(y_n | \beta_n) p_\beta(\beta_n | \alpha), \quad (1)$$

where a top-level latent variable α ties together all the β_n variables (de Finetti, 1961; Wang and Blei, 2015).¹ Reweighted probabilistic models broadens this approach by facing new challenges, including missing modeling assumptions, misspecified nonlinearities, and skewed data.

The second theme is robust Bayesian analysis, which studies sensitivity with respect to the prior (Berger et al., 1994). Recent advances directly focus on sensitivity of the posterior (Minsker et al., 2014; Miller and Dunson, 2015), or the posterior predictive distribution (Kucukelbir and Blei, 2015). We draw connections to these as we develop RPMS.

2 Reweighted Probabilistic Models

Reweighted probabilistic models (RPMS) offer a new approach to robust modeling. The idea is to automatically identify observations and hidden patterns that match the assumptions of the model.

2.1 Definitions

An RPM scaffolds over a probabilistic model, $p_\beta(\beta) \prod_n \ell(y_n | \beta)$. Raise each likelihood to a latent weight and posit a prior on the weights. This gives the joint density

$$p(y, \beta, w) = p_\beta(\beta) p_w(w) \prod_{n=1}^N \ell(y_n | \beta)^{w_n}. \quad (2)$$

This approach applies to models with likelihoods that factorize over the observations. (We discuss non-exchangeable models in Section 5.) Figure 1 depicts an RPM as a graphical model. Specific models may have additional structure, such as a separation of local and global latent variables (Hoffman et al., 2013), and fixed parameters; we ignore these in the graphical models.

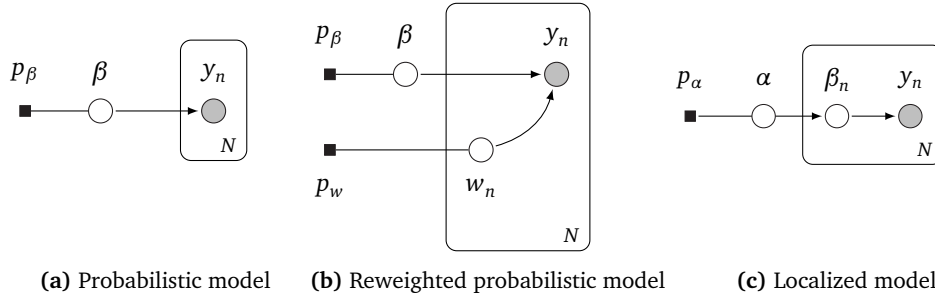


Figure 1: RPMS begin with a probabilistic model (a) and introduce a set of weights w as latent variables. This gives an RPM (b) that explores which data observations match its assumptions. (c) Localization, instead, builds a hierarchical model. (The Appendix shows when a localized model can also be described as a RPM.)

The reweighted model introduces a set of weights; these are latent variables, each with support $w_n \in \mathbb{R}_{>0}$. To gain intuition, consider how these weights affect terms in the likelihood. A weight w_n that is close to zero flattens out its corresponding likelihood $\ell(y_n | \beta)^{w_n}$; a weight that is larger than one makes its likelihood more peaked. The prior $p_w(w)$ can ensure that not too many likelihood terms get flattened; in this sense, it plays an important regularization role.

We study three options for this prior: a bank of Beta distributions, a scaled Dirichlet distribution, and a bank of Gamma distributions.

¹ Localization also relates to James-Stein shrinkage; Efron (2010) connects these dots.

Bank of Beta priors. This option constrains each weight to live $w_n \in (0, 1)$. We posit an independent prior for each weight

$$p_w(w) = \prod_{n=1}^N \text{Beta}(w_n; a, b) \quad (3)$$

and use the same parameters a and b for all weights. This is the most conservative option for the RPM; capping all weights to be less than one ensures that none of the likelihoods ever becomes more peaked than it was in the original model.

The parameters a, b offer an expressive language to describe prior belief on the weights. For example, setting both parameters less than one makes the Beta act like a “two spikes and a slab” prior, encouraging weights to be close to zero or one, but not in between. As another example, setting a greater than b encourages weights to lean towards one.

Scaled Dirichlet prior. This option ensures the sum of the weights equals N . We posit a symmetric Dirichlet prior on all the weights

$$\begin{aligned} w &= N\nu \\ p_\nu(\nu) &= \text{Dirichlet}(a\mathbf{1}) \end{aligned} \quad (4)$$

where a is a scalar parameter and $\mathbf{1}$ is a $(N \times 1)$ vector of ones. If all the weights were one, then the sum of the weights would be N . The Dirichlet option maintains this balance; while certain likelihoods may become more peaked, others will flatten to compensate.

The concentration parameter a is an intuitive way to capture prior beliefs. Small values for a allow the model to easily up- or down-weight many data observations; larger values for a prefer a smoother distribution of weights. The Dirichlet option connects to the bootstrap approach in (Kucukelbir and Blei, 2015), which also preserves the sum of weights as N .

Bank of Gamma priors. This option does not constrain the weights. We posit an independent prior for each weight

$$p_w(w) = \prod_{n=1}^N \text{Gamma}(w_n; a, b) \quad (5)$$

and use the same parameters a and b for all weights. We do not recommend this option, as observations can be arbitrarily up- or down-weighted. In this paper, we only consider Equation (5) for theoretical analysis.

The bank of Beta and Dirichlet options perform similarly. We prefer the Beta option as it is more conservative, yet find the Dirichlet to be less sensitive to its parameters. The Beta prior performs best when its shape parameter a scales with the dataset size N . We set $a \approx 0.1$ for small N , and $a \approx 100$ for large N . We explore these options in Section 3.

2.2 Theory and intuition

How can theory justify RPMs? Here we investigate the robustness properties of the RPM. These analyses intend to give intuition; the Appendix presents these results in full technical detail.

Intuition under a Gamma prior. Consider an iid prior for the weights. Then, the logarithm of the RPM joint density is

$$\log p(y, \beta, w) = \sum_{n=1}^N \log p_w(w_n) + \log p_\beta(\beta) + \sum_{n=1}^N w_n \log \ell(y_n | \beta).$$

We compute the maximum-a-posterior (MAP) estimate of the weights w . The partial derivative is

$$\frac{\partial \log p(y, \beta, w)}{\partial w_n} = \frac{d \log p_w(w_n)}{dw_n} + \log \ell(y_n | \beta) \quad \text{for all } n = 1, \dots, N. \quad (6)$$

Plug the Gamma prior (eq. (5)) into the partial derivative (eq. (6)) and set it equal to zero. This gives the MAP estimate of w_n as

$$\hat{w}_n = \frac{a - 1}{b - \log \ell(y_n | \beta)}. \quad (7)$$

The MAP estimate \hat{w}_n is an increasing function of the log likelihood of y_n . This reveals that \hat{w}_n shrinks the contribution of observations that are unlikely under the log likelihood; in turn, this encourages the MAP estimate for $\hat{\beta}$ to describe the majority of the observations. Thus, when the MAP estimate $\hat{\beta}$ is close to the true β , \hat{w}_n down-weights corrupted observations that do not belong to the true data generating process. This is how an RPM makes a probabilistic model more robust.

A similar argument holds for other priors on weights. We formalize this intuition and generalize it in the following theorem.

Theorem 1 (informal) *Denote the true value of β as β_0 . Let the posterior mean of β under the weighted and unweighted model be $\hat{\beta}_w$ and $\hat{\beta}_u$ respectively. Assume mild conditions on p_w , ℓ and the corruption level, and that $|\ell(y_n | \hat{\beta}_w) - \ell(y_n | \beta_0)| < \epsilon$ holds $\forall n$ with high probability. Then, there exists an N^* such that for $N > N^*$, we have $|\hat{\beta}_u - \beta_0| \succeq_2 |\hat{\beta}_w - \beta_0|$, where \succeq_2 denotes second order stochastic dominance. (Details in the Appendix.)*

The theorem shows that an RPM improves the posterior estimate of the latent variable β . How much of an improvement does it give? We can quantify this through the influence function (IF) of $\hat{\beta}_w$.

Consider a statistic T that takes as input data from some distribution F . The $\text{IF}(z; T, F)$ measures how much an additional observation valued at z affects the statistic T . Define

$$\text{IF}(z; T, F) = \lim_{t \rightarrow 0^+} \frac{T(t\delta_z + (1-t)F) - T(F)}{t}$$

for z where this limit exists. Roughly, the IF measures the asymptotic bias on $T(F)$ caused by an observation that does not come from F . We consider a statistic T to be robust if its IF is a bounded function of z (Huber, 2011).

Say a value z has log likelihood $\log \ell(z, \beta_0)$ that is nearly $-\infty$; we think of this z as an outlier. Consider the weight function induced by p_w . This is a function of the log likelihood $w(\log \ell(\cdot, \beta_0))$, like in Equation (7).

Theorem 2 *If $\lim_{a \rightarrow -\infty} w(a) = 0$ and $\lim_{a \rightarrow -\infty} a \cdot w(a) < \infty$, then*

$$\text{IF}(z; \hat{\beta}_w, \ell(\cdot | \beta_0)) \rightarrow 0, \text{ as } \ell(z | \beta_0) \rightarrow 0.$$

This result shows that an RPM is robust, in that its IF goes to zero for arbitrarily unlikely measurements. This is true for all three prior options we present. (Details in the Appendix.)

2.3 Inference and computation

We now turn to inferring the posterior of an RPM, $p(\beta, w | y)$. The posterior lacks an analytic closed-form expression for all but the simplest of models; even if the original model admits such a posterior for β , the reweighted posterior may take a different form.

To approximate the posterior, we appeal to probabilistic programming. A probabilistic programming system enables a user to write a probability model as a computer program and then compile that program into an inference executable. Automated inference is the backbone of such systems: it inputs a probability model, expressed as a program, and outputs an efficient algorithm for inference.

We use three automated inference algorithms offered in Stan, a probabilistic programming system (Carpenter et al., 2015). Stan provides MAP estimation through limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimization (Nocedal and Wright, 2006); variational inference through automatic differentiation variational inference (ADVI) (Kucukelbir et al., 2015); and Hamiltonian Monte Carlo simulation through the No-U-Turn sampler (NUTS) (Hoffman and Gelman, 2014).

3 Empirical Study

We study RPMS under four types of model mismatch. This section involves simulations of realistic scenarios; the next section presents a recommendation system example using real data. We default to NUTS (Hoffman and Gelman, 2014) for inference in all experiments, unless otherwise stated. The computational cost of inferring the weights is unnoticeable.

3.1 Corrupted observations: a network wait-time example

A router receives packets over a network and measures the time it waits for each packet. Say we typically observe wait-times that follow a Poisson distribution with rate $\beta = 5$. We model each measurement using a Poisson likelihood $\ell(y_n | \beta) = \text{Poisson}(\beta)$ and posit a Gamma prior on the rate as $p_\beta(\beta) = \text{Gam}(a = 2, b = 0.5)$.

Imagine that $F\%$ percent of the time, the network fails. During these failures, the wait-times come from a Poisson with much higher rate $\beta = 50$. Thus, the data actually contains a mixture of two Poisson distributions; yet, our model only assumes one. (Details in the Appendix.)

How do we expect an RPM to behave in this situation? Say the network failed 25% of the time. Figure 2a shows the posterior distribution on the rate β . The original posterior is centered at 18; this is troubling, not only because the rate is wrong, but also because of how confident the posterior fit is. Localization introduces greater uncertainty in its inference, yet still estimates a rate around 15. The RPM correctly identifies that the majority of the observations come from $\beta = 5$, which also coincides with its prior assumption on the rate. Observations from when the network failed are down-weighted, which gives a confident posterior centered at five.

Figure 2b shows posterior 95% credible intervals of β under failure rates up to $F = 45\%$. The RPM is robust to corrupted measurements that defy its assumptions. The model is insensitive to the prior on the weights; both Beta and Dirichlet options perform similarly. From here on, we continue our focus on the Beta option. The results in this simple simulation study directly follow the intuition and theory from Section 2; we now move on to more challenging forms of model mismatch.

3.2 Missing latent groups: predicting color blindness

Color blindness is unevenly hereditary: it is much higher for men than for women (Boron and Boulpaep, 2012). Say we are not aware of this fact. We have a dataset of both genders with each individual’s color blindness status and his/her relevant family history. Consider analyzing this data using logistic regression. It can only capture one hereditary group. As the ratio of

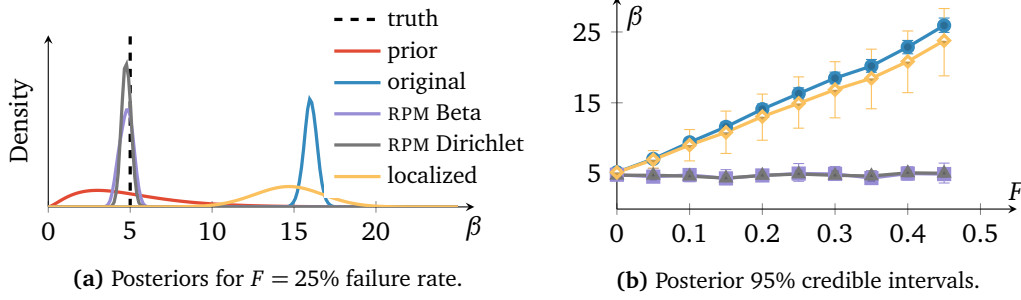


Figure 2: Corrupted observations simulation study. We compare Beta(0.1, 0.01) and Dir(1) as priors for the reweighted probabilistic model. **(a)** Posterior distributions on β show a marked difference in detecting the correct wait-time rate. **(b)** Posterior 95% confidence intervals across failure rates show consistent behavior for both Beta and Dirichlet priors. ($N = 100$ with 50 replications.)

females to males increase in the data, logistic regression misrepresents both groups. Yet an RPM can mitigate the missing group effect and focus on the dominant male hereditary trait.

We simulate this scenario by drawing binary indicators of color blindness $y_n \sim \text{Bernoulli}(1/(1 + \exp(-p_n)))$ where the p_i 's come from two latent groups: men exhibit a stronger dependency on family history ($p_n = 0.5x_n$) than women ($p_n = 0.01x_n$). We simulate family history as $x_n \sim \text{Unif}(-10, 10)$. Consider a Bayesian logistic regression model without intercept. Posit a prior on the slope as $p_\beta(\beta) = \mathcal{N}(0, 10)$ and assume a Beta(0.1, 0.01) prior on the weights. (Details in the Appendix.)

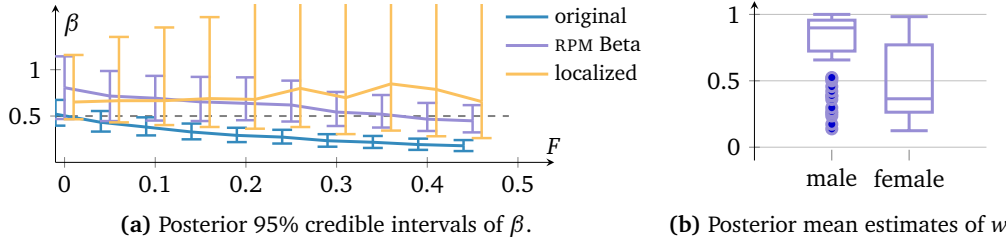


Figure 3: Missing latent groups study. **(a)** Credible intervals for the reweighted and localized model always include the true β , as we vary the percentage of females in the data. Dataset size $N = 100$ with 50 replications. **(b)** The reweighted model down-weights all of the females; the majority of the males receive a weight close to one. ($F = 25\%$.)

Figure 3a shows the posterior 95% credible intervals of β as we vary the percentage of females from $F = 0\%$ to 45% . A horizontal line indicates the correct slope for men, $\beta = 0.5$. As the size of the missing latent group (women) increases, the original model quickly shifts its credible interval away from 0.5. The reweighted and localized models both contain 0.5 for all percentages, yet the localized model exhibits much higher variance in its estimates. Figure 3b shows how the reweighted model identifies and down-weights females in the dataset; men are identified correctly as the major group and receive weights close to one.

3.3 Covariate dependence misspecification: a lung cancer risk study

Consider a study of lung cancer risk. While tobacco usage exhibits a clear connection, other factors may also contribute. For instance, obesity and tobacco usage appear to interact, with evidence towards a quadratic dependence on obesity (Odegaard et al., 2010).

Denote tobacco usage as x_1 and obesity as x_2 . We study three models of lung cancer risk dependency on these covariates. We are primarily interested in understanding the effect of tobacco usage; thus we focus on β_1 , the regression coefficient for tobacco. In each model, some form of covariance misspecification discriminates the true structure from the assumed structure.

For each model, we simulate a dataset of size $N = 100$ by randomly simulated covariates $x_1 \sim \mathcal{N}(10, 5^2)$ and $x_2 \sim \mathcal{N}(0, 10^2)$ and regression coefficients $\beta_{0,1,2,3} \sim \text{Unif}(-10, 10)$. Consider a Bayesian linear regression model with prior $p_\beta(\beta) = \mathcal{N}(0, 10)$. (Details in the Appendix.)

| true structure | model structure | original mean(std) | RPM Beta mean(std) | localization mean(std) |
|---|---------------------------------------|-----------------------|-----------------------|---------------------------|
| $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ | $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ | 3.16(1.37) | 2.20 (1.25) | 2.63(1.85) |
| $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2$ | $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ | 30.79(2.60) | 16.32 (1.96) | 21.08(5.20) |
| $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ | $\beta_0 + \beta_1 x_1$ | 0.58(0.38) | 0.60(0.40) | 0.98(0.54) |

Table 1: Absolute deviations of posterior mean β_1 estimates. (50 replications.)

Table 1 summarizes the misspecification and shows absolute differences on the estimated β_1 regression coefficient. The RPM yields better absolute deviations of β_1 estimates in the first two models. These highlight how the RPM leverages datapoints useful for estimating β_1 . The third model is particularly challenging because obesity is completely ignored in the misspecified model. Here, the RPM gives similar results to the original model; this highlights that RPMs can only use information that is available. Since the model lacks any dependence on x_2 , the RPM cannot compensate for this.

3.4 Skewed data: component selection in a mixture model

The Gaussian mixture model (GMM) is a versatile model for density estimation and clustering (Bishop, 2006; Murphy, 2012). While real data may indeed come from a finite mixture of clusters, there is no reason to assume each cluster is distributed as a Gaussian. Inspired by the

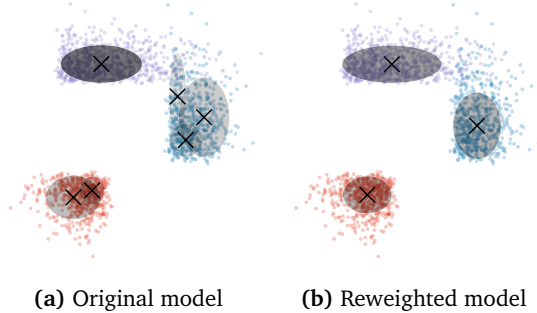


Figure 4: Fitting a 30 mixtures of Gaussian model skew normal data. The shade of each cluster indicates the inferred mixture proportions ($N = 2000$). experiments in Miller and Dunson (2015), we show how a reweighted GMM reliably recovers the correct number of components in a mixture of skewnormals dataset.

The GMM requires specifying the (unknown) number of clusters K . One way to navigate this is to fit a model with large K and posit a sparse Dirichlet prior on the mixture proportions. We simulate three clusters from two-dimensional skewnormal distributions and fit a GMM with $K = 30$. Here we use ADVI, as NUTS struggles with mixture model inference (Kucukelbir et al., 2016). (Details in the Appendix.)

Figure 4 shows posterior mean estimates from the original GMM; it finds six clusters. In contrast, the RPM correctly identifies three clusters. Datapoints in the tails of each cluster get down-weighted; these are datapoints that do not match the Gaussianity assumption of the model.

4 Case Study: Poisson factorization for recommendation

We now turn to a study of real data: a recommendation system. One form of recommendation data comes as a binary matrix of users (of a video streaming service) and the movies they watch. How can we identify patterns from such data? Poisson factorization (PF) offers a flexible solution (Cemgil, 2009). The idea is to infer a K -dimensional latent space of user preferences θ and movie attributes β . The inner product $\theta^\top \beta$ determines the rate of a Poisson likelihood for each binary measurement; Gamma priors on θ and β promote sparse patterns. As a result, PF finds interpretable groupings of movies, often clustered according to popularity or genre. (Full model in the Appendix.)

How does classical PF compare to its reweighted counterpart? As input, we use a subset of the MovieLens 1M dataset with $M = 1000$ randomly sampled movies. We place sparse iid Gamma(0.001, 1) priors on the preferences and attributes. We can reweight users or items. We focus on users and place a Beta(50, 1) prior on their weights. For this model, we use ADVI. (Localization is computationally infeasible with PF; it would require a separate “copy” of θ for each movie, along with a separate β for each user. This dramatically increases computational cost.)

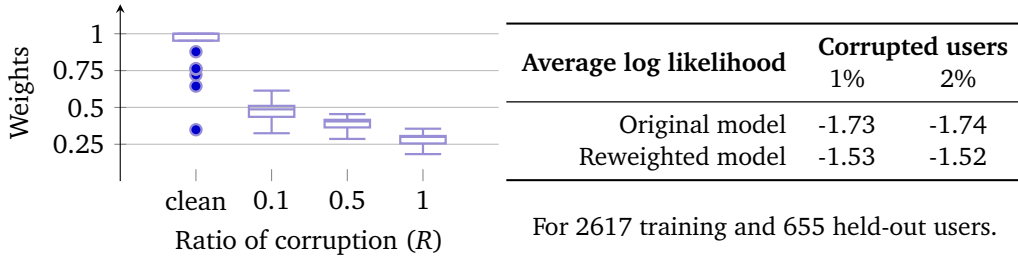


Figure 5: Weights for corrupted users and corresponding average held-out log likelihood results.

We begin by analyzing the clean dataset. (In the next paragraph, we introduce corruption.) Reweighting improves the average held-out log likelihood from -1.68 of the original model to -1.53 of the corresponding RPM. The boxplot in Figure 5 shows the inferred weights. The majority of users receive weight close to one, but a few users are down-weighted. These are film enthusiasts who appear to indiscriminately watch many movies from many genres. (The Appendix shows an example.) These users do not contribute towards identifying movies that go together; this explains why the reweighted model down-weights them and improves predictive accuracy.

Corrupted accounts. Recall the example from our introduction. A child typically watches popular animated films, but her parents occasionally use her account to watch a horror film. We simulate this by corrupting a small percentage of users. We replace a ratio $R = (0.1, 0.5, 1)$ of these users’ movies with randomly selected movies.

The boxplot in Figure 5 shows the weights we infer for these corrupted users, based on how many of their movies we randomly replace. The weights get lower as we corrupt more of their movies. The table shows how this leads to higher held-out predictive accuracy; down-weighting these corrupted users leads to better prediction.

5 Discussion

Probabilistic models embody assumptions about data. However, these assumptions may not always hold in every dataset we analyze. Reweighted probabilistic models offer a systematic approach to mitigating various forms of mismatch. The idea is to raise each data likelihood to a weight; inferring the weights along with the hidden patterns down-weights problematic datapoints. We demonstrate how this introduces robustness across four types of model mismatch.

There are several avenues for development. One direction is to extend RPMS to non-exchangeable data, such as time series. Many time series models admit exchangeable likelihood approximations (Guinness and Stein, 2013). In other cases, a non-overlapping windowing approach would also work. Another idea is to connect the weights to measures of goodness-of-fit; the sum of the weights gives an indication of model mismatch and relates to information criteria.

Acknowledgments

We thank Adji Dieng and Yuanjun Gao for their insightful comments. This work is supported by NSF IIS-1247664, ONR N00014-11-1-0651, and DARPA FA8750-14-2-0009.

References

- Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betró, B., et al. (1994). An overview of robust Bayesian analysis. *Test*, 3(1):5–124.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer New York.
- Boron, W. F. and Boulpaep, E. L. (2012). *Medical Physiology*. Elsevier.
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2015). Stan: a probabilistic programming language. *Journal of Statistical Software*.
- Cemgil, A. T. (2009). Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009.
- de Finetti, B. (1961). The Bayesian approach to the rejection of outliers. In *Proceedings of the Fourth Berkeley Symposium on Probability and Statistics*.
- Efron, B. (2010). *Large-Scale Inference*. Cambridge University Press.
- Feng, J., Xu, H., Mannor, S., and Yan, S. (2014). Robust logistic regression and classification. In *NIPS*.
- Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459.
- Guinness, J. and Stein, M. L. (2013). Transformation to approximate independence for locally stationary Gaussian processes. *Journal of Time Series Analysis*, 34(5):574–590.
- Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.
- Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler. *JMLR*, 15(1):1593–1623.
- Huber, P. J. (2011). *Robust statistics*. Springer.
- Insua, D. R. and Ruggeri, F. (2012). *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media.
- Kucukelbir, A. and Blei, D. M. (2015). Population empirical Bayes. In *UAI*.
- Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. M. (2015). Automatic variational inference in Stan. *NIPS*, pages 568–576.
- Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2016). Automatic differentiation variational inference. *arXiv preprint arXiv:1603.00788*.
- Mansinghka, V., Selsam, D., and Perov, Y. (2014). Venture: a higher-order probabilistic programming platform with programmable inference. *arXiv:1404.0099*.
- McWilliams, B., Krummenacher, G., Lucic, M., and Buhmann, J. M. (2014). Fast and robust least squares estimation in corrupted linear models. In *NIPS*.
- Miller, J. W. and Dunson, D. B. (2015). Robust Bayesian inference via coarsening. *arXiv preprint arXiv:1506.06101*.
- Minsker, S., Srivastava, S., Lin, L., and Dunson, D. B. (2014). Robust and scalable Bayes via a median of subset posterior measures. *arXiv preprint arXiv:1403.2660*.
- Murphy, K. P. (2012). *Machine Learning: a Probabilistic Perspective*. MIT Press.
- Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer.

- Odegaard, A. O., Pereira, M. A., Koh, W.-P., Gross, M. D., Duval, S., Mimi, C. Y., and Yuan, J.-M. (2010). BMI, all-cause and cause-specific mortality in Chinese Singaporean men and women. *PLoS One*, 5(11).
- Provost, F. and Fawcett, T. (2001). Robust classification for imprecise environments. *Machine learning*, 42(3):203–231.
- Shafieezadeh-Abadeh, S., Esfahani, P. M., and Kuhn, D. (2015). Distributionally robust logistic regression. In *NIPS*.
- Song, Q., Hu, W., and Xie, W. (2002). Robust support vector machine with bullet hole image classification. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 32(4):440–448.
- Wang, C. and Blei, D. M. (2015). A general method for robust Bayesian modeling. *arXiv preprint arXiv:1510.05078*.
- Yu, Y., Aslan, O., and Schuurmans, D. (2012). A polynomial-time form of robust regression. In *NIPS*.